REMARKS

In response to the Office Action mailed September 3, 2009, the Assignee respectfully requests reconsideration. Claims 1, 4-5, 7-11, 14-15, 17-23 and 29-35 were previously pending for examination. Claims 1, 4, 7-11, 14, 17-23 and 29-35 have been amended herein. No claims have been canceled or added. As a result, claims 1, 4-5, 7-11, 14-15, 17-23 and 29-35 remain pending, with claims 1, 11 and 21 being independent. No new matter has been added.

Rejections Under 35 U.S.C. 103

The Office Action rejects each of the independent claims under 35 U.S.C. 103(a) as purportedly being obvious over Lumelsky (U.S. Patent No. 6,081,780) in view of Applicant's Admitted Prior Art (AAPA). The Office Action rejects each of the dependent claims under 35 U.S.C. 103(a) as purportedly being obvious over Lumelsky and AAPA, either alone or in combination with Saon et al ("Maximum Likelihood Discriminant Feature Spaces," 2000, IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, 5-9 June 2000, pages 1129-1132). The Assignee respectfully traverses each of these rejections.

I. Overview of Some Embodiments

The present application describes systems and methods for converting a text input to synthesized speech in a manner that mimics the style and pronunciation of a spoken example of the text input (page 1, lines 6-10). When a user supplies a spoken example of a text string to be text-to-speech synthesized, some embodiments can extract prosodic parameters from the spoken example, and use those prosodic parameters in generating the synthetic speech waveform (page 8, lines 1-8). Prosodic parameters are speech quality attributes, such as pitch, duration, energy values, etc. for multiple speech segments in the spoken audio signal (page 1, lines 21-42; page 10, lines 23-24).

Some embodiments allow a user to input a text string and speak an example of a desired pronunciation of the text string (page 9, lines 9-12). A prosody analyzer can then process the

spoken audio signal to extract prosodic parameters from it, e.g., pitch and energy contours (i.e., sets of pitch and energy values extracted from various time indexes during the audio signal) (page 10, line 20 – page 11, line 1). An alignment module can extract the duration of each unit of the text (e.g., word or phoneme) as produced in the spoken example (page 13, lines 18-21). This can be accomplished by aligning the audio signal with the text.

The specification describes using a Viterbi algorithm as one technique for aligning the audio signal with the text (page 13, lines 14-17; page 14, lines 12-13). The Viterbi algorithm itself was known in the art and had been used in various applications, not limited to speech-related fields. For example, Forney (G.D. Forney, Jr., "The Viterbi Algorithm", Proc. IEEE, vol. 61, pp. 268-278, 1973, incorporated in the specification by reference) describes how the algorithm was first developed for decoding convolutional codes, for example in space communications fields. Forney also describes applications of the Viterbi algorithm to intersymbol interference, continuous-phase frequency-shift keying, and text recognition. The embodiment described in the present application that uses the Viterbi algorithm describes using it in a new way to extract durational parameters from aligning an input audio signal with input text. These durational parameters may then be used, along with prosodic parameters extracted from the audio signal, in a text-to-speech (speech synthesis) application.

After prosodic and durational parameters have been extracted by the prosody analyzer, a conversion module can process the parameters to generate an input for a text-to-speech (TTS) engine (page 15, line 24 – page 16, line 4). The input may be in a format such as SSML (Speech Synthesis Markup Language), which provides the text string with prosody markup that can be processed by the TTS engine to produce a synthetic speech waveform with the specified prosodic parameters (page 16, line 9 – page 17, line 7). The duration parameters extracted by the alignment module allow the prosodic parameters from various time indexes during the audio signal to be mapped to the appropriate text units when the text is converted to synthetic speech (page 16, lines 5-8). The TTS input so generated may then be used by the TTS engine in converting the text string to synthetic speech (page 18, lines 9-11).

The foregoing overview is provided solely for the convenience of the Examiner. It should be appreciated that each of the claims may not be limited in the manner described in the overview above. Therefore, the Examiner is requested not to rely upon the overview above for determining whether each of the claims distinguishes over the art of record, but to do so based solely upon the language of the claims themselves and the arguments presented below.

II. Overview of Lumelsky

Lumelsky describes adjusting a phonetic representation of a text used in synthesizing that text to speech. The adjustment is made through corrective feedback based on spectral comparison of an audio signal produced by the TTS system and an audio signal spoken by a human narrator (Lumelsky: col. 9, lines 1-6). When the system receives a text to be synthesized, the system generates initial prosodic parameters with reference to an internal dictionary, rather than with reference to a spoken example (Lumelsky: col. 12, line 44 – col. 13, line 8). A first synthetic speech audio signal is then generated using the initial prosodic parameters, and spectrally compared to a spoken audio signal from the narrator (Lumelsky: col. 13, lines 42-58). The spectral distance between the synthetic audio signal and the narrator's audio signal then determines the amount of correction to be applied to the TTS system's prosodic parameters in the next synthesis iteration (Lumelsky: col. 14, lines 17-25). Synthesis and correction iterations are repeated until satisfactory results are achieved (Lumelsky: col. 14, lines 25-27).

III. Independent Claim 1

The Assignee respectfully submits that the Office Action's rejection of independent claim 1 as purportedly being obvious over Lumelsky and AAPA is improper, as the Office Action does not state how Lumelsky could be modified in view of AAPA, the Office Action's interpretation of what is admitted prior art according to the specification is factually incorrect, and claim 1 distinguishes over any combination of Lumelsky and AAPA.

Claim 1 as amended recites, "extracting duration parameters by aligning the audio signal with the text string", and "automatically generating at least one text-to-speech (TTS) input using the

prosodic parameters and the duration parameters". Neither Lumelsky nor AAPA meets these limitations.

At page 3, the Office Action concedes that "Lumelsky does not specifically teach an alignment process for aligning the spoken utterance with a corresponding text string." Lumelsky therefore necessarily also fails to disclose or suggest "extracting duration parameters by aligning the audio signal with the text string" and "automatically generating at least one text-to-speech (TTS) input using the prosodic parameters and the duration parameters", as required by claim 1.

The Office Action asserts that "aligning a spoken utterance with a corresponding text string was well known in the art. Applicant's admitted prior art (AAPA) specifically indicates implementation of Viterbi alignment was well known in the art." However, as discussed above, while the present application mentions that the Viterbi algorithm was known, it does not say that it was known to use the Viterbi algorithm for aligning an audio signal with text, let alone for using such alignment for "extracting duration parameters" to be used in "automatically generating at least one text-to-speech (TTS) input". The Office Action's interpretation of AAPA is thus factually incorrect.

There is simply no support for the Office Action's assertion at page 4 that "applying the known standard techniques of automatic alignment of speech to text to the TTS system of Lumelsky would have yielded predictable results and resulted in an improved system." The Office Action does not even allege what modifications to Lumelsky it is believed one skilled in the art would have made based on knowledge of the Viterbi algorithm, and as such the assertion that making some unspecified change to Lumelsky could yield "predictable results" is entirely unsupported. In addition, the Office Action's assertion that any "predictable results" could be obtained from using the Viterbi algorithm for "automatic alignment of speech to text" in the system of Lumelsky is based upon the factually incorrect assertion that AAPA discloses "implementation of Viterbi alignment" for "aligning a spoken utterance with a corresponding text string". As discussed above, the specification does not state that such use of the Viterbi algorithm was known in the art.

Even if one of skill in the art were somehow motivated to align speech to text within the system of Lumelsky, any such alignment would have no use under Lumelsky's teachings, and any alleged resulting system would be far from "improved", as asserted by the Office Action. As discussed above, Lumelsky generates an initial synthesis of an input text using default system parameters, then compares the resulting synthesized audio with spoken audio from a narrator. Text alignments of any form would be irrelevant to the system's described function of comparing synthesized audio to spoken audio for corrective feedback, and the system would presumably simply ignore any outputs of alignment rather than using them in the manner recited in claim 1.

Even if the Viterbi algorithm of AAPA were combined with the system of Lumelsky, the alleged combination would fail to meet all limitations of claim 1. Integrating knowledge of the Viterbi algorithm into the system of Lumelsky would simply result in an extra step of performing the Viterbi algorithm on something unspecified, for no apparent purpose. Neither Lumelsky nor AAPA discloses a use of an alignment technique for "extracting duration parameters" or "automatically generating at least one text-to-speech (TTS) input using... the duration parameters", as required by claim 1.

For at least these reasons, claim 1 patentably distinguishes over Lumelsky and AAPA, and it is respectfully requested that the rejection of claim 1 be withdrawn.

Claims 4-5, 7-10, 29-30 and 33 depend from claim 1 and are allowable for at least the same reasons. Accordingly, it is respectfully requested that the rejections of these claims over Lumelsky and AAPA, or over Lumelsky, AAPA and Saon, be withdrawn.

IV. Independent Claim 11

Independent claim 11 as amended recites, "extracting duration parameters by aligning the audio signal with the text string", and "automatically generating at least one text-to-speech (TTS) input using the prosodic parameters and the duration parameters". For reasons that should be clear from the foregoing discussion of Lumelsky and AAPA, these references, whether alone or in combination, fail to disclose or suggest the above limitations of claim 11. Therefore, claim 11

patentably distinguishes over Lumelsky and AAPA, and it is respectfully requested that the rejection of claim 11 be withdrawn.

Claims 14-15, 17-20, 31-32 and 34 depend from claim 11 and are allowable for at least the same reasons. Accordingly, it is respectfully requested that the rejections of these claims over Lumelsky and AAPA, or over Lumelsky, AAPA and Saon, be withdrawn.

V. Independent Claim 21

Independent claim 21 as amended recites, "an alignment module for extracting duration parameters by aligning the input text string with the audio signal, and a conversion module for generating the at least one TTS system input using the prosodic parameters and the duration parameters". For reasons that should be clear from the foregoing discussion of Lumelsky and AAPA, these references, whether alone or in combination, fail to disclose or suggest the above limitations of claim 21. Therefore, claim 21 patentably distinguishes over Lumelsky and AAPA, and it is respectfully requested that the rejection of claim 21 be withdrawn.

Claims 22-23 and 35 depend from claim 21 and are allowable for at least the same reasons. Accordingly, it is respectfully requested that the rejections of these claims be withdrawn.

General Comments on Dependent Claims

Because each of the dependent claims depends from a base claim that is believed to be in condition for allowance, the Assignee believes that it is unnecessary at this time to argue the further distinguishing features of all of the dependent claims. However, the Assignee does not necessarily concur with the interpretation of the dependent claims as set forth in the Office Action, nor does the Assignee concur that the basis for the rejection of any of the dependent claims is proper. Therefore, the Assignee reserves the right to specifically address in the future the further patentability of the dependent claims not specifically addressed herein.

CONCLUSION

In view of the foregoing, the present application is believed to be in condition for allowance. A Notice of Allowance is respectfully requested. The Examiner is requested to call the undersigned at the telephone number listed below if this communication does not place the application in condition for allowance.

If this response is not considered timely filed and if a request for an extension of time is otherwise absent, any necessary extension of time is hereby requested. If there is a fee occasioned by this response, including an extension fee, the director is hereby authorized to charge any deficiency or credit any overpayment in the fees filed, asserted to be filed or which should have been filed herewith to our Deposit Account No. 23/2825, under Docket No. N0484.70760US00.

Dated: December 3, 2009

Respectfully submitted,
Nuance Communications, Inc.

By _____
Richard F. Giunta
Registration No.: 36,149
WOLF, GREENFIELD & SACKS, P.C.
Federal Reserve Plaza
600 Atlantic Avenue
Boston, Massachusetts 02210-2206
617.646.8000

x12/03/2009x